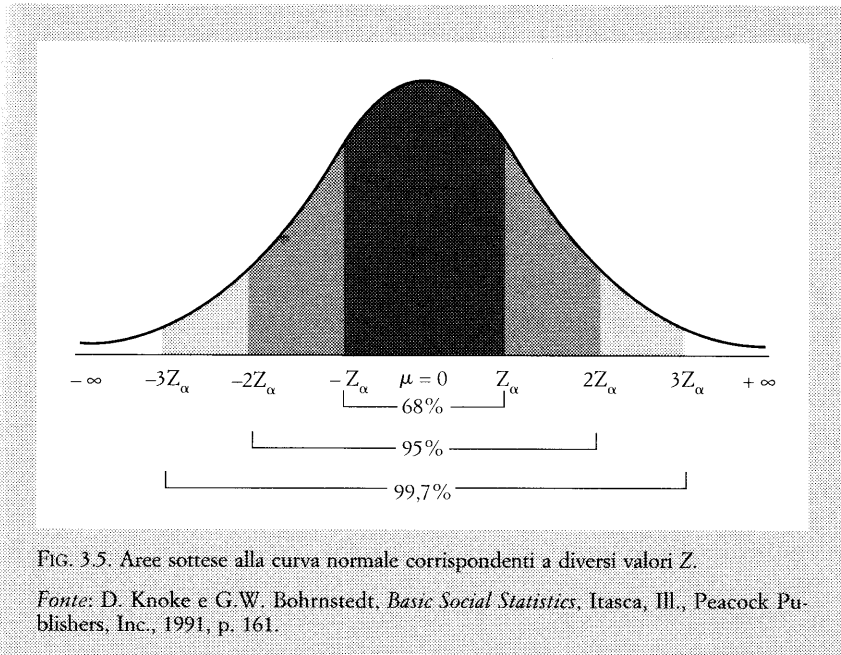




stabilire che $0,5 - 0,025 = 0,475$. Quindi, utilizzando la solita tabella B1 riportata nell'appendice B, possiamo vedere che il valore $Z_{\alpha/2}$ associato a 0,475 è pari a 1,96; naturalmente, poiché la distribuzione normale è simmetrica, si ha che $-Z_{\alpha/2} = -1,96$. Dunque, il 95% dell'area sottesa alla curva normale standardizzata si colloca fra $-Z_{\alpha/2} = -1,96$ e $Z_{\alpha/2} = 1,96$.

5. Il teorema del limite centrale

Un uso molto importante della curva normale si basa sul **teorema del limite centrale**, che afferma quanto segue:

Se tutti i possibili campioni casuali di numerosità N vengono estratti da una data popolazione avente media μ_Y e varianza σ_Y^2 , all'aumentare di N le medie di questi campioni approssimeranno una distribuzione normale, con media μ_Y e varianza σ_Y^2/N . Formalmente:

$$\mu_{\bar{Y}} = \mu_Y$$

$$\sigma_{\bar{Y}}^2 = \sigma_Y^2/N$$

Il teorema del limite centrale, in sostanza, afferma che data una certa numerosità campionaria N , la media di tutte le medie campiona-

rie possibili è uguale alla media della popolazione dalla quale i campioni sono stati estratti in modo casuale. Inoltre, la varianza di questa nuova ipotetica distribuzione è uguale a $1/N$ della varianza della popolazione. Non è necessario formulare alcun assunto sulla forma della distribuzione.

La distribuzione ipotetica di tutte le possibili medie di campioni di numerosità N è chiamata **distribuzione campionaria delle medie campionarie**. Una distribuzione campionaria di medie è formata dalle medie di tutti i campioni di numerosità N che sarebbe possibile estrarre da una data popolazione. Poiché qualsiasi popolazione di una certa dimensione contiene miliardi e miliardi di possibili campioni differenti, nessuno può veramente calcolare le medie che formano una distribuzione campionaria. Questa, cioè, rappresenta un costrutto puramente teorico. Tuttavia, poiché il teorema del limite centrale mette in relazione due parametri della popolazione (μ_Y e σ_Y^2) con la media e la varianza della distribuzione campionaria, la forma di quest'ultima è completamente determinata da questi due soli parametri. Il teorema del limite centrale garantisce che la media campionaria approssima sempre più la media della popolazione mano a mano che la dimensione del campione (N) aumenta, in quanto la varianza della distribuzione campionaria diminuisce all'aumentare di N . La deviazione standard di una distribuzione campionaria è detta **errore standard**. Questo equivale alla radice quadrata della varianza della distribuzione campionaria:

$$\sigma_{\bar{Y}} = \frac{\sigma_Y}{\sqrt{N}}$$

Sapendo che le medie campionarie seguono una distribuzione normale, indipendentemente dalla popolazione dalla quale i corrispondenti campioni sono stati estratti, e assumendo che N sia grande, possiamo giungere a importanti conclusioni. Supponiamo di estrarre un campione casuale di numerosità pari a 400 osservazioni da una popolazione con $\mu_Y = 100$ e $\sigma_Y = 15$. L'errore standard della distribuzione campionaria delle medie, in questo caso, può essere calcolato come segue:

$$\sigma_{\bar{Y}} = \frac{15}{\sqrt{400}} = 0,75$$

Ciò che sappiamo della curva normale ci dice che il 95% di tutte le medie campionarie sono comprese in un intervallo di $\pm 1,96$ errori standard intorno alla media della popolazione. Pertanto, in questo esempio il 95% delle medie campionarie rientrano in un intervallo compreso fra 98,53 e 101,47, cioè nell'intervallo $100 \pm (1,96)(0,75)$. Il teorema del limite centrale ci dice che a) la media di una distribuzione

campionaria di medie è uguale alla media della popolazione (in questo caso 100), e *b*) solo il 5% di tutte le medie campionarie possibili non rientrano nell'intervallo compreso fra 98,53 e 101,47. Supponiamo ora di aumentare la dimensione del campione da 400 a 1.000 osservazioni. La conseguenza di ciò è che l'errore standard diventa ancora più piccolo, cioè:

$$\sigma_{\bar{Y}} = \frac{15}{\sqrt{1.000}} = 0,47$$

Pertanto, con la nuova numerosità campionaria il 95% delle medie campionarie saranno comprese fra 99,08 e 100,92, cioè nell'intervallo $100 \pm (1,96)(0,47)$. Dunque, complessivamente solo una media campionaria su venti non rientrerà in questo intervallo. Possiamo allora avere fiducia nel fatto che *qualsiasi* campione formato da 1.000 osservazioni ci darà una stima accurata della media della popolazione dalla quale è stato estratto.

Sebbene il teorema del limite centrale possa essere applicato solo a campioni di numerosità sufficientemente ampia, non è possibile stabilire precisamente quando un campione è «sufficientemente» grande. Alcuni manuali suggeriscono un limite inferiore di 30 osservazioni, altri indicano una numerosità minima pari a 100. Sulla base dell'esperienza, possiamo dire che quando un campione è formato da almeno 100 casi, la distribuzione campionaria delle medie approssima in modo soddisfacente una distribuzione normale. Per campioni formati da 30 o meno casi, al contrario, è difficile assumere una distribuzione campionaria normale. Per i campioni di numerosità compresa fra 30 e 100, si può assumere con cautela che il teorema del limite centrale sia applicabile, a meno che nella popolazione di riferimento il fenomeno di interesse abbia una distribuzione dalla forma molto irregolare.

6. Stime campionarie puntuali e intervalli di confidenza

Il teorema del limite centrale possiede un importante corollario: la media di un campione casuale ($\bar{Y}$) rappresenta la migliore stima singola della media della popolazione (μ_Y) dalla quale il campione è stato estratto. La media campionaria è una **stima puntuale** della media della popolazione, perché consiste in un solo valore stimato. Possiamo anche costruire un **intervallo di confidenza** intorno a questa stima, che ci consente di esprimere la misura in cui riteniamo che questo intervallo contenga la vera media della popolazione.

Anche in questo caso il teorema del limite centrale ci viene in aiuto, permettendoci di usare la curva normale per costruire un intervallo di confidenza caratterizzato da una specifica probabilità α . La formula

per calcolare il limite inferiore e quello superiore di un intervallo di confidenza di livello α è la seguente:

$$\bar{Y} \pm (Z_{\alpha/2})(\sigma_{\bar{Y}})$$

Per calcolare il **limite superiore di confidenza** (LSC) bisogna aggiungere alla media campionaria il prodotto del valore critico ($Z_{\alpha/2}$) moltiplicato per l'errore standard della media campionaria. Per calcolare, invece, il **limite inferiore di confidenza** (LIC) bisogna sottrarre questo prodotto dalla media campionaria. Per esempio, se scegliamo un livello $\alpha = 0,05$, i valori critici $Z_{\alpha/2}$ che ci permettono di collocare il 2,5% dell'area sottesa alla curva normale in ciascuna delle due code della distribuzione normale stessa sono $-1,96$ e $1,96$. Possiamo allora aspettarci che il 95% degli intervalli compresi fra $\pm 1,96$ errori standard dalla media campionaria contengano la media della popolazione μ_Y . Per un intervallo di confidenza del 95% il limite inferiore di confidenza è pari a $\bar{Y} - 1,96\sigma_{\bar{Y}}$, mentre il limite superiore di confidenza è uguale a $\bar{Y} + 1,96\sigma_{\bar{Y}}$. Analogamente, per un intervallo di confidenza del 99% il limite inferiore di confidenza è pari a $\bar{Y} - 2,58\sigma_{\bar{Y}}$, mentre il limite superiore di confidenza è uguale a $\bar{Y} + 2,58\sigma_{\bar{Y}}$. Come si può vedere, minore è il livello α , più ampio è l'intervallo di confidenza.

Il simbolo $Z_{\alpha/2}$, introdotto nel paragrafo 3.4, può essere ora definito con maggiore precisione. Nella tavola B1 riportata nell'appendice B, $Z_{\alpha/2}$ denota il valore che delimita un'area della coda destra della curva normale pari a $\alpha/2$. Se vogliamo essere «fiduciosi» al 95% che un certo intervallo contenga μ_Y , allora ne consegue necessariamente che $\alpha = 0,05$ e $\alpha/2 = 0,025$.

Gli intervalli di confidenza devono essere interpretati con una certa cautela. La tentazione più forte è quella di concludere che un dato intervallo ha il 95% di probabilità di contenere la media della popolazione. Di fatto, però, una volta che un dato intervallo è stato costruito, la probabilità che esso contenga la media della popolazione può essere solo pari a 1 o a zero, a seconda che la media della popolazione sia inclusa o meno nell'intervallo. Dunque, l'unica affermazione che possiamo fare è che se costruiamo molti intervalli di confidenza dello stesso tipo, il 95% di questi intervalli includerebbe la media della popolazione.

La figura 3.6 illustra il concetto di intervallo di confidenza. La linea verticale rappresenta la vera media della popolazione, una costante pari – in questo esempio – a 50,5. Le linee orizzontali, invece, rappresentano diversi intervalli di confidenza costruiti intorno alle medie di 15 campioni casuali estratti dalla stessa popolazione. Come si può vedere, tutti gli intervalli tranne due (il quinto e il quattordicesimo dall'alto) includono la media della popolazione μ_Y . Le stime puntuali (cioè le medie campionarie) sono riportate accanto a ciascun intervallo. Se potessimo costruire gli intervalli di confidenza di tutti i possibili

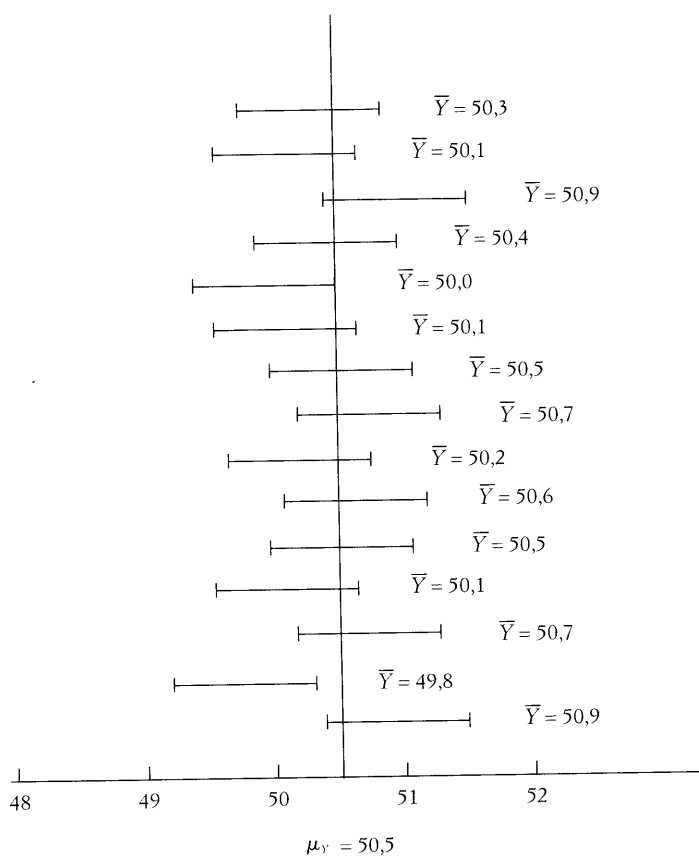


FIG. 3.6. Illustrazione del concetto di intervallo di confidenza.

Fonte: D. Knoke e G.W. Bohrnstedt, *Basic Social Statistics*, Itasca, Ill., Peacock Publishers, Inc., 1991, p. 167.

campioni casuali di numerosità N , troveremmo che $1 - \alpha$ di essi includono la media della popolazione, mentre i rimanenti α la escludono.

In generale, per un dato intervallo di confidenza, dimensioni campionarie maggiori producono intervalli più ristretti. Supponiamo di sapere che $\sigma_Y = 15$ e di calcolare $\bar{Y} = 51$ per un campione casuale di numerosità pari a 100. Dati questi valori, un intervallo di confidenza del 95% avrà un limite inferiore pari a $51 - 1,96(15/\sqrt{100}) = 48,06$ e un limite superiore uguale a $51 + 1,96(15/\sqrt{100}) = 53,94$. Supponiamo ora di passare a un campione di numerosità pari a 500 casi, lasciando invariati gli altri valori. In questo caso l'intervallo di confidenza del

95% avrà un limite inferiore pari a $51 - 1,96(15/\sqrt{500}) = 49,69$ e un limite superiore uguale a $51 + 1,96(15/\sqrt{500}) = 52,31$. Per un dato livello α , dunque, l'ampiezza dell'intervallo di confidenza diminuisce all'aumentare della numerosità N del campione.

Per costruire un intervallo di confidenza intorno a una media campionaria è necessario che siano soddisfatte tre condizioni: 1) il campione deve essere stato estratto casualmente; 2) N deve essere «sufficientemente» grande (almeno maggiore di 30); 3) la varianza della popolazione (σ_Y^2) deve essere nota, in modo tale che l'errore standard $\sigma_{\bar{Y}}$ possa essere calcolato. Nella maggior parte dei casi il terzo requisito rimane insoddisfatto. Chiaramente, se conoscessimo σ_Y^2 non avremmo bisogno di analizzare i dati di un campione. Quando N è sufficientemente grande, il problema può essere risolto; in questi casi, infatti, una buona stima dell'errore standard della distribuzione campionaria può essere ottenuta utilizzando la deviazione standard del campione, cioè s_Y . Un accento circonflesso posto sopra $\sigma_{\bar{Y}}$ indica che si tratta di un valore stimato:

$$\hat{\sigma}_{\bar{Y}} = \frac{s_Y}{\sqrt{N}}$$

I dati riportati nella tabella 2.3 danno una stima campionaria degli anni medi di istruzione formale pari a 12,9, con una deviazione standard pari a 2,98. Il campione su cui si basano queste statistiche ha una dimensione uguale a 1.510 casi. Inserendo tali statistiche campionarie nella formula appena mostrata possiamo ottenere la seguente stima dell'errore standard:

$$\hat{\sigma}_{\bar{Y}} = \frac{2,98}{\sqrt{1.510}} = 2,98/38,86 = 0,077$$

Dunque, possiamo essere «fiduciosi» al 95% che i veri anni medi di istruzione formale sono compresi nell'intervallo $12,9 \pm (1,96)(0,077)$, cioè fra 12,75 e 13,05 anni.